# WASPER: A BULGARIAN-LANGUAGE MODEL FOR DETECTING PROPAGANDA IN SOCIAL MEDIA CONTENT

**Asst. Prof. Dr. Yolina Petrova, PhD**
*New Bulgarian University, Identrics*

**Todor Kiryakov, Devora Kotseva, Boryana Kostadinova**
*Identrics*

***Abstract:***

*This paper introduces WASPer, a classification model designed to detect propaganda in Bulgarian-language social media content. In response to the rising threat of AI-generated disinformation and the regulatory requirements of the EU's Digital Services Act (DSA), WASPer aims to provide a practical and scalable solution for identifying manipulative narratives online. A thematically diverse dataset was constructed by combining manually annotated organic content and synthetic examples generated with a Bulgarian language model (BgGPT). Each text was human-annotated based on the presence of rhetorical techniques commonly associated with propaganda. The dataset was used to train WASPer (a fine-tuned version of the BgGPT 7B Instruct v0.2 model), achieving an F1 score of 0.853 on the test set. WASPer supports the detection of harmful or misleading content in digital spaces such as comment sections and social media threads, contributing to efforts to meet DSA obligations for transparency and risk mitigation.*

**Keywords**: propaganda detection, digital service act, social media, artificial intelligence

## 1. Introduction

Large Language Models (LLMs), such as Llama, Mistral, GPT-4[1] and their successors, represent a significant leap in artificial intelligence (AI), capable of generating text that is virtually indistinguishable from human writing. These models, trained on vast datasets from the internet, have a deep understanding

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

of language patterns, enabling them to produce coherent, contextually relevant, and persuasive narratives. While this technology has numerous beneficial applications, from automating customer service to leveraging content creation, it also poses significant risks when misused.

In fact, one of the most concerning instances of misuse of LLMs is their potential to generate propaganda on a massive scale. Synthetically generated content can be created rapidly and in large volumes, with minimal oversight. LLMs can tailor messages to specific audiences, mimic individual writing styles, and even generate responses in real time. This turns them into a particularly effective tool for shaping public opinion in a subtle and efficient manner, turning AI-generated propaganda into a critical challenge for democracy and national security. This issue, a key aspect of „digital warfare," threatens the integrity of both traditional and social media.

Sophisticated troll networks and automated bots are increasingly being deployed to manipulate online discourse and shape public perception. A recent example involves Russian operatives allegedly using nearly 1,000 fake AI-generated accounts on the social media platform *X* to impersonate Americans and spread propaganda, which highlights the real-world impact of such tools[2]. This case is just a single illustration of how AI technologies are being leveraged to amplify propaganda efforts, enabling the large-scale creation and dissemination of disinformation. Moreover, recent research[3] reveals that between January 1, 2022, and May 1, 2023, the presence of AI-generated news articles on mainstream websites increased by 55.4%, while on sites well-known for spreading misinformation, it surged by 457%.

As these tactics evolve and AI-driven content grows rapidly, the need for effective methods to identify AI-driven propaganda becomes increasingly urgent. Ensuring the credibility of information across digital platforms is essential to preserve public discourse. The growing threat of (AI-enabled) disinformation has not gone unnoticed by regulators. The Digital Services Act (DSA), adopted by the European Union, seeks to create a safer digital environment by imposing legal obligations on platforms to address illegal content, increase algorithmic transparency, and mitigate systemic risks to public discourse. In particular, the DSA mandates that very large online platforms implement measures to counter disinformation and protect the democratic process. This regulatory framework highlights the urgent need for technical solutions that can support compliance efforts, including the detection of AI-generated propaganda. In the digital age, where information spreads at an unprecedented speed, the ability to identify and counteract propaganda is more crucial than ever – particularly when it is generated or amplified by sophisticated AI systems.

[2] Pequeño IV, A. 2024, *Russia Impersonated Americans Using Nearly 1,000 Fake AI-Generated X Accounts*, Feds Allege.

[3] Hanley, H. & Durumeric, Z. (2023). Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. arXiv 2305.09820.

The work of the Identrics team addresses this challenge through the development of WASPer, a model designed to detect propaganda in Bulgarian social media content. By focusing on a non-English language context, the approach fills a significant gap in the existing literature and aligns with the DSA's broader goal of safeguarding the integrity of online communication across diverse linguistic and regional landscapes.

### 1.1. What is propaganda and why focus on it?

According to the often-quoted definition by the Institute for Propaganda Analysis, *„propaganda is the expression of opinions or actions carried out deliberately by individuals or groups with a view to influence the opinions or actions of other individuals or groups for predetermined ends through psychological manipulations.*"[4] In his seminal work on the topic, Jacques Ellul[5] further expands the concept, noting that propaganda encompasses various forms of psychological influence – including psychological action (efforts to „modify opinions by purely psychological means"), psychological warfare (attempts to destroy an adversary's morale), re-education and brainwashing (actions to transform „an adversary into an ally"), and public and human relations (efforts „to adapt the individual into a society, to a living standard, to an activity").

Ellul also distinguishes between two major types of propaganda: *political* and *sociological*[6]. Political propaganda is relatively easy to identify, as it is closely related to the domain of politics - such as election campaigns, referendums, protests, and armed conflicts. In contrast, sociological propaganda is less direct but permeates different aspects of our daily lives, including our digital surroundings. It permeates everyday life – including our digital environments – and works by integrating individuals into dominant social norms and belief systems. As Ellul observes, *„nothing is easier than to graft a direct propaganda onto a setting prepared by sociological propaganda."*[7] In other words, sociological propaganda lays the groundwork for political propaganda by shaping the underlying narratives and belief systems through which people interpret events – even when those events initially appear apolitical.

In a more recent influential study, Jowett and O'Donnell[8] define propaganda as *„the deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behaviour to achieve a response that furthers the desired intent of the propagandist"*. They emphasize the calculated nature of propaganda – it *„is carefully thought out ahead of time to select what will be the most effective strategy*

[4] Lee, A. & Lee, E. (1937) *The Fine Art of Propaganda: A Study of Father Coughlin's Speeches*. Institute for Propaganda Analysis [online].

[5] Ellul, J. (1973) *Propaganda. The Formation of Men's Attitudes*. New York: Vintage Books.

[6] Ibid.

[7] Ibid.

[8] Jowett, G. & O'Donnell, V. (1999). *Propaganda and Persuasion*. Thousand Oaks, CA: Sage.

*to promote an ideology and maintain an advantageous position“*[9]. Emotional appeal and logical fallacy are central tools: propaganda exploits emotional triggers, cognitive biases, and reasoning errors to suppress critical thinking and encourage audiences to adopt simplified, polarized interpretations of complex issues.

Importantly, the rhetorical strategies commonly referred to as „propaganda techniques“ are not exclusive to propagandistic discourse. Techniques such as exaggeration, oversimplification, whataboutism, or stereotyping also appear in everyday human communication. What transforms these into instruments of propaganda is their deliberate use to influence public opinion toward a predetermined, often political, objective. Therefore, the presence of propaganda techniques in a text should be seen as a strong indication but not absolute evidence that the text is propagandistic, as their classification ultimately depends on the communicator’s intent and the broader communicative context[10].

### 1.2. Why focus on social media data?

Social media platforms – and increasingly, the comment sections of traditional media – are among the primary venues for public discourse and information exchange today. These spaces are particularly vulnerable to the spread of propaganda, due to the unmoderated nature of user-generated content and the speed at which information can be shared and amplified. The rapid advancement and democratization of large language models (LLMs), combined with the low barrier to entry for participating in online discussions, have significantly increased the risk of synthetic propaganda infiltrating public conversations. WASPer is therefore designed to focus specifically on social media data and related digital environments, where propaganda can spread quickly and where detection tools are most urgently needed.

## 2. Defining Propaganda for Binary Classification - Training dataset

To support a binary classification task – determining whether a given text contains propaganda – we constructed a labelled dataset sourced from Bulgarian social media platforms and online news comment sections. Trained annotators labelled each example as Propaganda or No Propaganda, depending on the presence of identifiable persuasive or manipulative techniques.

The labelling process was informed by a comprehensive review of established propaganda theory. One of the earliest systematic classifications comes from *The Fine Art of Propaganda*[11], published by the Institute for Propaganda Analysis,

[9] Jowett, G. & O‘Donnell, V. (1999). *Propaganda and Persuasion*. Thousand Oaks, CA: Sage.

[10] Jowett, G., & O’Donnell, V. (2018). *Propaganda & persuasion.* (Seventh edition). SAGE.

[11] Lee, A. & Lee, E. (1937) *The Fine Art of Propaganda: A Study of Father Coughlin’s Speeches*. Institute for Propaganda Analysis [online].

which introduced seven classic methods such as *Name-calling*, *Glittering Generalities*, and *Bandwagon*. Later contributions by Silverman[12], Torok[13], and Piskorski et al.[14] extended this body of work by introducing additional, context-sensitive techniques. Drawing from these studies, a reference list of 32 textual propaganda techniques was compiled. These were used not for multi-label annotation, but to guide binary decisions: if a text exhibited *any* of these techniques, it was labelled Propaganda. If none were present, the text was labelled No Propaganda. To ensure relevance in text-only environments, the criteria for inclusion in this reference list were as follows: 1) Techniques must be recognisable using only textual content, without the need for visual, auditory, or behavioural context; 2) Recognition must be possible without relying on external fact-checking or verifying the truth value of the message; and 3) Techniques must be detectable within standalone documents (e.g., individual comments or posts), without requiring patterns across multiple messages.

To ensure thematic diversity, the dataset includes content from the comment sections of four major Bulgarian online news platforms, as well as publicly available social media posts. Collection was focused on topics frequently associated with online propaganda, including:

- *Domestic politics* – e.g., Bulgarian politicians, attitudes towards Russia and the West;
- *International politics* – e.g., US politics and elections, conspiracy theories;
- *Military conflict* – e.g., Ukraine/Russia, the Middle East;
- *Environment* – e.g., the Green Deal, wind turbines, solar energy;
- *Society*, *social conflict*s – e.g. LGBT, gender, the Istanbul convention;
- *Science* – e.g., vaccines, food;
- *Artificial intelligence* – e.g., deep fakes.

This thematic coverage helps ensure the model's generalization ability and mitigates the risk of topic-specific overfitting. The dataset comprises both organic (naturally occurring) and synthetic examples. All organic texts were manually reviewed and labelled by trained annotators following the binary criteria. Annotators consulted the 32-technique reference list to assess whether texts employed manipulative rhetorical strategies but did not label specific techniques.

To supplement the naturally occurring propaganda data, we generated synthetic examples using the BgGPT language model[15]. These examples were created through few-shot prompting, using organic propaganda samples as seeds. Prompts

[12] Silverman, H. (2011) *Reuters: Principles Of Trust Or Propaganda?* Journal of Applied Business Research; Laramie Vol. 27, Iss. 6, 93-115.

[13] Torok, R. (2015) *Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming.* In Proceedings of the Australian Security and Intelligence Conference, ASIC '15, pages 58-65.

[14] Piskorski, J. et al. (2023) *News Categorization, Framing and Persuasion Techniques: Annotation Guidelines*, European Commission, Ispra, JRC132862.

[15] Alexandrov, A., Raychev, V., Müller, M. N., Zhang, C., Vechev, M., & Toutanova, K. (2024). *Mitigating Catastrophic Forgetting in Language Transfer via Model Merging. arXiv preprint arXiv:2407.08699.*

were designed to elicit outputs corresponding to different persuasive styles and propaganda strategies. The generated content was manually reviewed and annotated using the same binary labelling criteria. This approach ensured sufficient representation of propaganda examples, particularly for underrepresented topics or styles. To avoid topic-based label leakage (e.g., associating certain topics only with propaganda), topic modelling was applied to balance the No Propaganda samples. Using BERTopic[16], we performed unsupervised clustering of text samples based on semantic similarity. This ensured that No Propaganda examples were drawn from the same thematic space as Propaganda examples. To guarantee that the non-propaganda dataset includes content related to the same topics as the propaganda examples, zero-shot topic modelling was also applied. In this approach, the topics described in Section 2 were used as a guide to locate non-propaganda examples discussing similar themes. This approach ensures that non-propaganda examples are present across the same range of topics as propaganda, allowing for balanced and comprehensive coverage of themes. This approach allowed us to balance the dataset effectively while minimizing manual annotation burden. We compiled a final training dataset of 734 examples. An evaluation set of 50 examples (25 *Propaganda*, 25 *No Propaganda*) was also constructed.

## 3. Binary Classification for Propaganda Detection

Binary classification is a machine learning approach where the objective is to assign each data instance to one of two predefined classes. In our case, the goal is to determine whether a given text is classified as Propaganda or No Propaganda. We used the previously described dataset and adopted the following data splits: 80% for training (587 examples), 10% for validation (73 examples), and 10% for testing (74 examples). During training, the model learns to distinguish between propaganda and non-propaganda based on the labelled examples. The validation set is used to tune hyperparameters and monitor overfitting. The test set is reserved for final evaluation on unseen data.

We fine-tuned BgGPT 7B Instruct v0.2[17], a model selected for its robust pretraining on Bulgarian-language data, which enhances its ability to capture the linguistic and contextual nuances relevant to this task. The performance of WASPer (the fine-tuned Propanda classification model) was assessed using three primary metrics: F1-micro (measures overall performance across all instances), F1-macro (gives equal weight to each class, regardless of class size), and F1-weighted (balances precision and recall while accounting for label distribution). The model demonstrated consistent improvements across all metrics over the course of training. After 11 epochs of training (on a single

[16] Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.*

[17] Alexandrov, A., Raychev, V., Müller, M. N., Zhang, C., Vechev, M., & Toutanova, K. (2024). *Mitigating Catastrophic Forgetting in Language Transfer via Model Merging. arXiv preprint arXiv:2407.08699.*

V100 GPU), it achieved a final training loss of 0.2335 and a final F1 score on the test set 0.853 (out of 1.00).

The resulting model is publicly available[18] to support further research, development, and operational use - particularly for social media platforms seeking to comply with the Digital Services Act (DSA) in the context of AI-generated disinformation and online manipulation.

## 4. Conclusions

The spread of propaganda, particularly when amplified by powerful AI tools, represents a significant challenge to the health of digital information ecosystems. With the growing accessibility of large language models (LLMs) and their proven capability to generate persuasive and realistic synthetic content, it becomes increasingly urgent to develop effective tools for detecting manipulative communication online. This paper addresses that need by introducing WASPer. WASPer was developed in response to both societal and regulatory imperatives. On one hand, the growing risk of AI-generated propaganda demands scalable, accurate detection mechanisms. On the other hand, legislative frameworks such as the European Union's Digital Services Act (DSA) impose explicit obligations on platforms to mitigate systemic risks related to disinformation and content manipulation. WASPer supports both objectives by offering a language-specific solution capable of identifying rhetorical manipulation in real-world digital content.

A key strength of our approach lies in the construction of a carefully curated dataset that combines organic and synthetic text examples across a wide thematic range, reflecting the complex and multifaceted nature of propaganda today. Guided by a theoretically grounded set of 32 rhetorical techniques, the binary labelling strategy enabled a clear and actionable classification task. We fine-tuned the BgGPT 7B Instruct v0.2 model on this dataset, leveraging its strong performance in Bulgarian language understanding. The resulting model achieved an F1 score of 0.853 on the held-out test set, demonstrating strong predictive performance across both classes. While promising, the model's performance is limited by the small size of the dataset and the subjective nature of manual annotations, which may not fully capture the diversity of real-world propaganda styles. Despite these constraints, the model shows potential not only for standalone classification tasks, but also for integration into real-time moderation workflows and content auditing pipelines.

The WASPer model and its methodology contribute to the body of tools aimed at empowering content moderation, platform accountability, and public discourse integrity. Although focused on Bulgarian, the approach can be generalized to other low-resource languages that currently lack tailored disinformation detection infrastructure. It also supports the broader EU policy objective of ensuring digital safety across linguistic and regional contexts, not just English-dominant environments.

[18] https://huggingface.co/identrics/wasper_propaganda_detection_bg

## BIBLIOGRAPHY

- Alexandrov, A., Raychev, V., Müller, M. N., Zhang, C., Vechev, M., & Toutanova, K. (2024). *Mitigating Catastrophic Forgetting in Language Transfer via Model Merging. arXiv preprint arXiv:2407.08699*.
- Ellul, J. (1973). *Propaganda. The Formation of Men's Attitudes*. New York: Vintage Books.
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794*.
- Hanley, H. & Durumeric, Z. (2023). Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. arXiv 2305.09820.
- Jiang, A. , Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D., Casas, D., ... & Sayed, W. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Jowett, G., & O'Donnell, V. (1999). *Propaganda and Persuasion*. Thousand Oaks, CA: Sage.
- Jowett, G., & O'Donnell, V. (2018). *Propaganda & persuasion* (Seventh edition.). SAGE.
- Lee, A. & Lee, E. (1937) *The Fine Art of Propaganda: A Study of Father Coughlin's Speeches*. Institute for Propaganda Analysis. Available at: https://archive.org/details/LeeFineArt [Accessed 8 Aug. 2024].
- Pequeño, A. (2024) *Russia Impersonated Americans Using Nearly 1,000 Fake AI-Generated X Accounts*, Available at: https://www.forbes.com/sites/antoniopequenoiv/2024/07/09/russia-impersonated-americans-using-nearly-1000-fake-ai-generated-x-accounts-feds-allege/ [Accessed 20 August 2024].
- Piskorski, J. et al. (2023) *News Categorization, Framing and Persuasion Techniques: Annotation Guidelines*, European Commission, Ispra, JRC132862.
- Silverman, H. (2011) *Reuters: Principles Of Trust Or Propaganda?* Journal of Applied Business Research; Laramie Vol. 27, Iss. 6, 93-115.
- Torok, R. (2015) *Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming*. In Proceedings of the Australian Security and Intelligence Conference, ASIC '15, pages 58-65, Perth, Australia.